
Journal of Biological Physics **29:** 341–362, 2003.

Protein Structure from X-Ray Diffraction

M.W. PARKER
Biota Structural Biology Laboratory, St. Vincent's Institute of Medical Research, 9 Princes Street, Fitzroy, Victoria 3065, Australia e-mail: mwp@rubens.its.unimelb.edu.au

Abstract. Protein crystallography is the study of the three-dimensional structures of proteins at near atomic resolution. It has provided a tremendous insight into the workings of numerous biological processes over the last few decades. The field has undergone a massive worldwide expansion over the last ten years, not only in academic laboratories, but also in the pharmaceutical industry. The main driving force for this expansion has been the promise of using three-dimensional atomic structures of proteins and other macromolecules to design lead drugs and to improve the action of existing drugs.

1. Introduction

The three-dimensional structures of proteins are essential for understanding protein function and activity. Detailed knowledge of protein structures has been vital for our understanding of numerous biological processes, from enzymatic reactions to immune evasion by viruses. In the case of enzymes their structures have revealed how substrates and inhibitors interact with them and has provided insight into the mechanisms of enzyme-catalyzed reactions. Similarly, recent structures of hormones, such as human growth hormone bound to its receptor, has formed the basis for understanding how signals are passed across cellular membranes.

The major technique for determining the atomic structures of proteins has been X-ray crystallography, although in the last decade nuclear magnetic resonance spectroscopy has proved a powerful tool for deciphering the atomic structures of small proteins ($<$ 25 kDa). The discovery of X-ray crystallography and its application for solving structures of molecules was made by Sir William Bragg and his son, Sir Lawrence Bragg. Some of Sir William's early work was performed in Australia when he was Professor of Natural History (Physics of the day) at Adelaide University. Sir Lawrence was born in Adelaide. Their achievements in X-ray crystallography were recognised with the award of the Nobel Prize in 1915. Protein crystallography can trace its origins back to 1934 when J.D. Bernal and Dorothy Crowfoot-Hodgkin at the Cavendish Laboratory in Cambridge (U.K.) discovered that crystals of the stomach protease pepsin yielded an X-ray diffraction pattern [1]. It was not until 1960 that the technical difficulties associated with deciphering an atomic structure from diffraction patterns of protein crystals were met and the first structure of a protein was published [2]. The first protein structures revealed that

the polypeptide chain of the protein folded into a well-defined three-dimensional structure using two major conformational elements: a coiled structure called a helix and an extended configuration termed a beta-strand. To some, these results were a surprise as it was predicted that proteins would have no defined structure based on studies of non-biological polymers. Obtaining protein structures in the early days was a very slow and painstaking effort with the result that less than a dozen protein structures had been determined by 1970. Advances in computing and molecular biology have dramatically speeded up the process with new protein structures now appearing at approximately 200 per month. At the time of writing there are over 16,500 protein structures available in the Protein Data Bank [3], the depository for models of macromolecular structures. The present collection contains a wide variety of macromolecules including enzymes, proteins bound to DNA, integral membrane proteins, antibody-antigen complexes and whole viruses. From its slow birth in the sixties, protein crystallography has now matured into an exciting and powerful technology that is still moving at an accelerating pace.

2. Crystallization

2.1. WHAT ARE CRYSTALS?

Crystals are three-dimensional periodic arrays of molecules. Unlike well-known crystals such as diamonds and sapphires, protein crystals have the special property that they require water as part of their structure, usually consisting about 50% of the crystal volume. If water is removed from protein crystals they lose their periodic order.

Crystals are required for X-ray diffraction experiments because scattering from individual molecules is far too weak to measure. Crystals act like an amplifier by increasing the scattering signal due to the multiple copies of molecules within them. Typical protein crystals are about 0.2 mm in size but usable crystals have been reported from tens of microns to a few millimeters.

The smallest repeating unit in a crystal is called the unit cell. There are 10^{14} such cells in a typical protein crystal with the contents of each unit cell being identical. The unit cells must pack well together with no packing defects if they are to be ordered enough to see useful diffraction. The unit cell is characterized by three sides, a, b and c and three angles α, β and γ (Figure 1). There are only seven crystal systems that are compatible with the building up of a regular crystal lattice. At one extreme, in the triclinic system, no sides are equal ($a \neq b \neq c$) and no angles are equivalent ($\alpha \neq \beta \neq \gamma$). At the other extreme, in the cubic system, all sides are equal ($a = b = c$) and all angles are equal to 90°.

Within the unit cell, the array of molecules may be described in terms of a number of possible space groups. A space group is one of 230 groups of symmetry operators (describing translation and rotation) that is consistent with an infinite array of molecules in the crystal. Examples of such symmetry operators are n-fold axes of symmetry (where n = 2, 3, 4 and 6) and mirror planes. Because amino

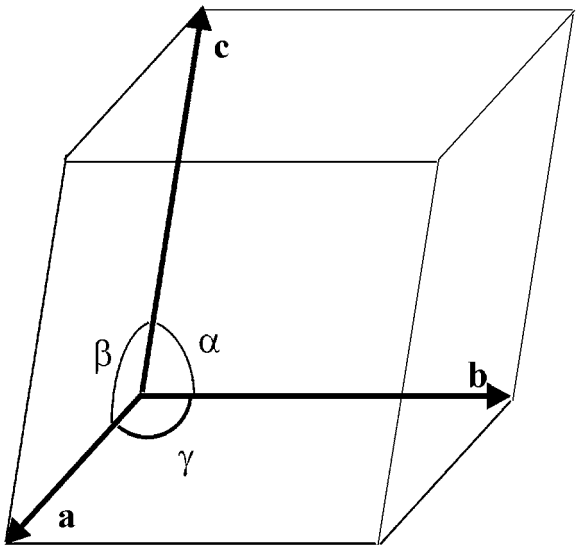


Figure 1. Definitions of unit cell parameters in a crystal (after G. Kong).

acids possess chirality (i.e. handedness) there are only 65 different space group possibilities for proteins as mirror planes and centers of inversion are not possible. The crystallographic symmetry within the unit cell can operate on one or more molecules that are collectively called the asymmetric unit of the crystal. In turn, where there is more than one molecule within the asymmetric unit, they might be related by local (or non-crystallographic) symmetry.

2.2. GROWING CRYSTALS

Proteins can be made to crystallize by the addition of certain precipitants such as salts and organic solvents, most commonly ammonium sulfate or polyethylene glycol, under usually precise conditions of pH, temperature and protein concentration. Protein crystallization is very complicated and considered a difficult art by its practitioners [4]. Many factors can influence successful crystallization including protein and precipitant concentrations, ionic strength, vibration, protein flexibility, protein purity, small molecule additives, temperature and so on. The detailed physics behind crystallization are not well understood. The process is usually considered in terms of phase diagrams where the vertical axis corresponds to the protein solubility and the horizontal axis refers to some experimental parameter such as pH or precipitant concentration (Figure 2). Consider the behavior of a typical protein solution. At low protein and precipitant concentrations the protein stays in solution (i.e. it is undersaturated). As the concentration of protein or precipitant increases the protein becomes less soluble until supersaturation occurs whereby the protein comes out of solution as either an aggregated mess (amorphous precipitate)

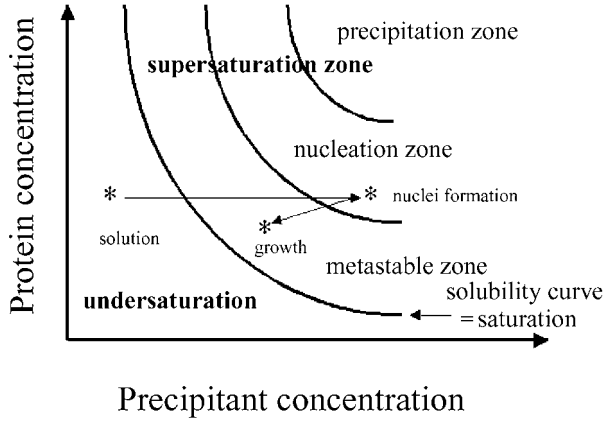


Figure 2. Phase diagram for crystallization.

or as ordered crystals. The zone in which crystals form is called the nucleation zone and the zone in which crystals grow is the metastable zone.

3. Diffraction Theory

3.1. INTRODUCTION

Small objects in the millimeter range are normally visualized using light microscopes where visible light, scattered from the object of interest, is collected and focused using the objective lens of the microscope. In order to visualize structures at the atomic scale it is necessary to work with electromagnetic radiation with wavelengths of the order of atomic bond distances (approximately 1 Å or 10^{-10} meters). X-rays have such suitable wavelengths. X-ray diffraction – the interference between waves scattered from individual atoms in a crystal – can be used to determine atomic structures. However, there are no lenses available to bend and focus the scattered X-rays. Instead atomic structures must be reconstructed using diffraction theory from the intensities of the diffracted waves which can be measured experimentally.

When X-rays impinge upon free electrons, the fluctuating electromagnetic field of the incident wave forces the electrons into oscillations of the same frequency as the incident wave. This oscillation results in the generation of secondary radiation of the same wavelength of the incident ray, but out of phase by 180°. This is called coherent or elastic scattering. Periodic waves can be defined by three parameters: the wavelength (λ), the amplitude and the phase (Figure 3).

The periodic wave can be expressed mathematically by:

$$\boldsymbol{F} = |F|\exp(i\phi) \tag{1}$$

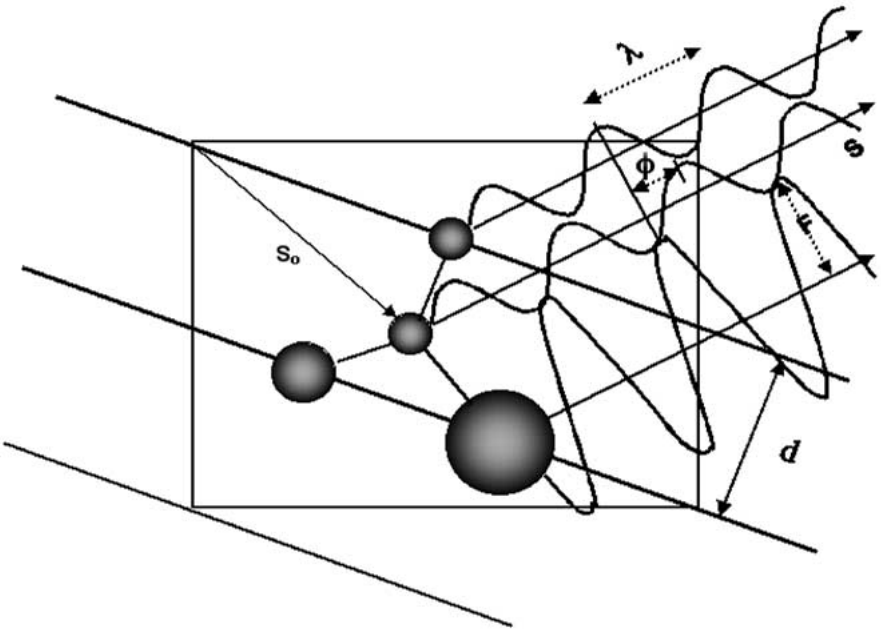


Figure 3. Diffraction from a molecule. A molecule is located in a two-dimensional unit cell. Each atom is shown by a sphere with diameter proportional to the number of electrons in the atom. X-rays are shone at the molecule in the direction indicated by the vector $\mathbf{s_o}$. X-ray waves are scattered from each atom. The distance between wave crests is the wavelength, λ, and the amplitude, F, is the difference between the peak height and the average displacement of the wave. The phase difference, ϕ, between two waves is shown. Imaginary Bragg planes (see below) are also indicated by the parallel straight lines separated by distance d.

Where the magnitude of $\boldsymbol{F}$ is the amplitude and ϕ is the phase with values between 0° and 360° (reflecting the periodic nature of waves).

Referring to Figure 4, the path difference between an X-ray scattered at some point P relative to that scattered by an electron at the origin is

$$\mathbf{r}.\mathbf{s_o} - \mathbf{r}.\mathbf{s} \tag{2}$$

where $\mathbf{r}$ is the vector distance of P from the origin and $\mathbf{s_o}$ and $\mathbf{s}$ represent the vectors of the incident and scattered rays respectively. If we choose the modulus of $\mathbf{s_o}$ and $\mathbf{s}$ to be $1/\lambda$, then the phase difference is given by $2\pi\,\mathbf{r}.\mathbf{S}$ where $\mathbf{S}$ is the vector difference between the incident and scattered waves. The vector $\mathbf{S}$ is called the scattering vector and is used to describe the position in diffraction space.

If we consider the lattice to be made up of atoms rather than electrons, we must consider the total wave scattered by the partial volume of the atom, $\mathbf{dv}$, and then sum up these individual contributions over the volume of the atom. The phase of a wave scattered scattered from $\mathbf{dv}$, relative to a point at a defined origin, depends critically on the position $\mathbf{r}$ as well as on the wave vector direction, $\mathbf{s}$, in relation to the incident wave vector direction, $\mathbf{s_o}$ (Figure 4). Thus diffraction from the scattered

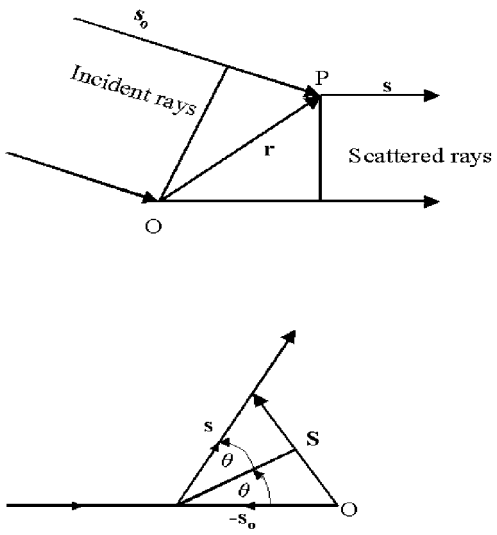


Figure 4. Scattering at a point P relative to an origin O and the relationship of the diffraction space vector **S** to the real space vectors $\mathbf{s_0}$ and **s**.

waves provides information about where atoms are located in the unit cell of a crystal, ie. diffraction depends on the atomic structure within the unit cell.

The total scattered wave, made up of scattering contributions from all the volume elements of the scattered object is given by the Fourier transform equation:

$$F(\mathbf{S}) = \int_V \int \rho(\mathbf{r})\exp(2\pi i\mathbf{r}.\mathbf{S})dv \tag{3}$$

where the amplitude is proportional to $\rho(\mathbf{r})$dv and the phase is $2\pi\,\mathbf{r}.\mathbf{S}$ The function $F(\mathbf{S})$ is referred to as the atomic scattering factor and is usually denoted by the symbol *f*. For electron-rich atoms the amplitude of the scattered wave is much greater than an atom with few surrounding electrons (Figure 3). The total wave scattered by a molecule of N atoms can be derived from vector addition of the atomic contributions:

$$\mathbf{G}(\mathbf{S}) = \sum_{\mathrm{j}=1,\mathrm{N}} f_j \exp(2\pi i\mathbf{r}_j.\mathbf{S}) \tag{4}$$

where $\mathbf{G}(\mathbf{S})$ is the molecular transform.

Diffraction patterns from protein crystals are characterized by diffraction maxima (referred to as spots or reflections) located on a periodic three dimensional grid (Figure 5). The location of any particular spot can be defined by three indices (*h*, *k* and *l*), sometimes referred to as the Miller indices of the diffraction spot. The origin of the grid is defined by the direction of the initial or primary beam, the majority of which passes through the crystal without being scattered. Because the distance between two adjacent spots in a row or column is inversely proportional to the unit cell dimensions, the diffraction patterns are commonly referred to as

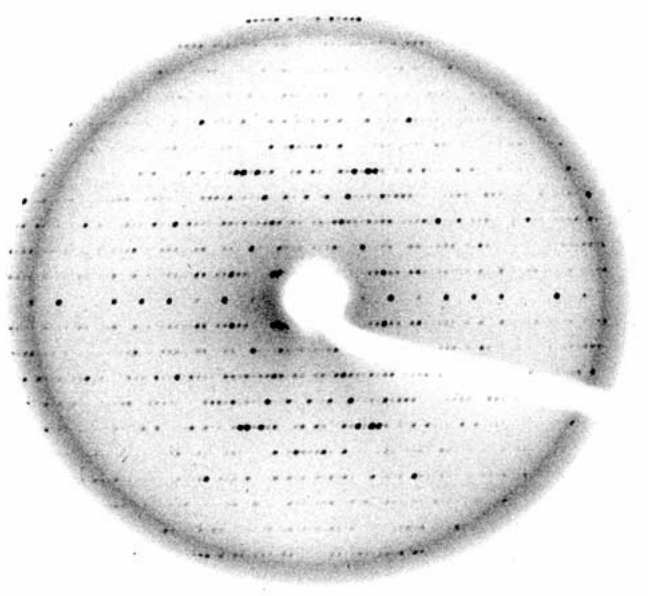


Figure 5. A diffraction pattern from a protein crystal. The direct (or primary) X-ray beam passes through the center of the picture and it defines the origin of the lattice. Because exposing the primary beam on a detector is not healthy for the instrument, it is routine to insert a piece of lead (the so-called backstop) between crystal and detector. The backstop and its holder are seen as the shadow in the picture. Each diffraction spot can be labeled with its own Miller indices, *h*, *k* and *l*. The symmetry observable in the diffraction pattern is due the space group of the crystals. The further the distance of spots from the center of the pattern, the higher the resolution.

'reciprocal space' as distinct from the real space of the electron density image that results from diffraction analysis.

If we define **r** in Figure 4 in terms of the unit cell vectors **a**, **b** and **c**, then the phase differences for a scattered beam of maximum intensity are

$$2\pi(\mathbf{a}.\mathbf{S}) = 2\pi h;\ 2\pi(\mathbf{b}.\mathbf{S}) = 2\pi k;\ 2\pi(\mathbf{c}.\mathbf{S}) = 2\pi 1 \tag{5}$$

where integers *h*, *k* and *l* are the Miller indices. The reciprocal lattice vector **S** can be given as:

$$\mathbf{S} = \mathrm{h}\mathbf{a}^* + \mathrm{k}\mathbf{b}^* + \mathrm{l}\mathbf{c}^* \tag{6}$$

where a*, b* and c* are the lattice constants of the diffraction lattice. We can obtain the scattering equation for the crystal by combining the equation expressing the fractional coordinates of the jth atom in terms of the lattice vectors:

$$\mathbf{r}_j = \mathbf{a}\mathrm{x}_j + \mathbf{b}\mathrm{y}_j + \mathbf{c}\mathrm{z}_j \tag{7}$$

with the equations (2d) so that

$$\mathbf{r}_j.\mathbf{S} = x_j.\mathbf{a}.\mathbf{S} + y_j.\mathbf{b}.\mathbf{S} + z_j.\mathbf{c}.\mathbf{S}$$
$$= hxj + kyj + lzj \quad (8)$$

If the volume of integration given in equation (3) corresponds to the unit cell of the crystal, the values obtained at integral values of **S**(h,k,l) are directly proportional to those of the whole crystal:

$$F_{h,k,l} = \sum_{j=1,natoms} f_j \exp(-B_j(\sin\theta/\lambda)^2_{h,k,l}).\ \exp\{2\pi i(hx_j + ky_j + lz_j)\} \quad (9)$$

where the summation is over all the atoms in the unit cell, each atom having an atomic scattering factor, *f*, and an atomic mobility factor, *B*, associated with it. The B-factor (sometimes referred to as the temperature factor) is related to atomic displacement, û, by the equation:

$$B = 8\pi^2\hat{u}^2 \quad (10)$$

Temperature factors have units of Å^2 with values typically less than 20 Å^2 in the core of the protein and values greater than 60 Å^2 on the protein surface.

The structure factor equation (9) represents the molecular transform sampled at each reciprocal lattice point. The structure factor is directly related to the experimentally determined intensity of each diffraction spot by:

$$I_{h,k,l} = C\,|F_{h,k,l}|^2 \quad (11)$$

where C is a proportionality factor that depends on various experimental factors.

3.2. BRAGG'S LAW

Bragg showed that if the diffraction from a crystal was considred as reflections from imaginary planes of atoms within the crystal then a equation could be formulated to predict where diffraction maxima would occur in a diffraction pattern. Consider a pair of parallel X-rays striking a pair of horizontal parallel planes as shown in Figure 6.

The parallel rays hit the planes in phase but the lower ray has a longer distance to travel than the upper one by the time they are both reflected. By simple trigonometry (see Figure 6) it can be shown that

$$n\lambda = 2d\sin\theta \quad (12)$$

The Bragg relationship shows that constructive interference of the waves will only occur when the path difference is some multiple of the wavelength, λ.

3.3. EWALD CONSTRUCTION

Another way of expressing Bragg's law is through the Ewald construction. In this construction a sphere with center at the crystal (C) of radius 1/λ is placed on a

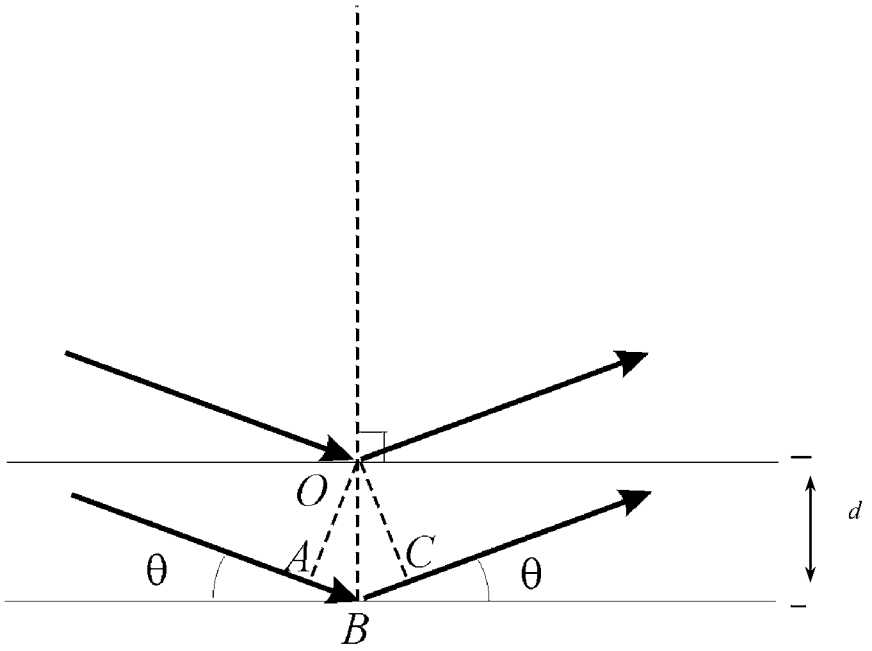


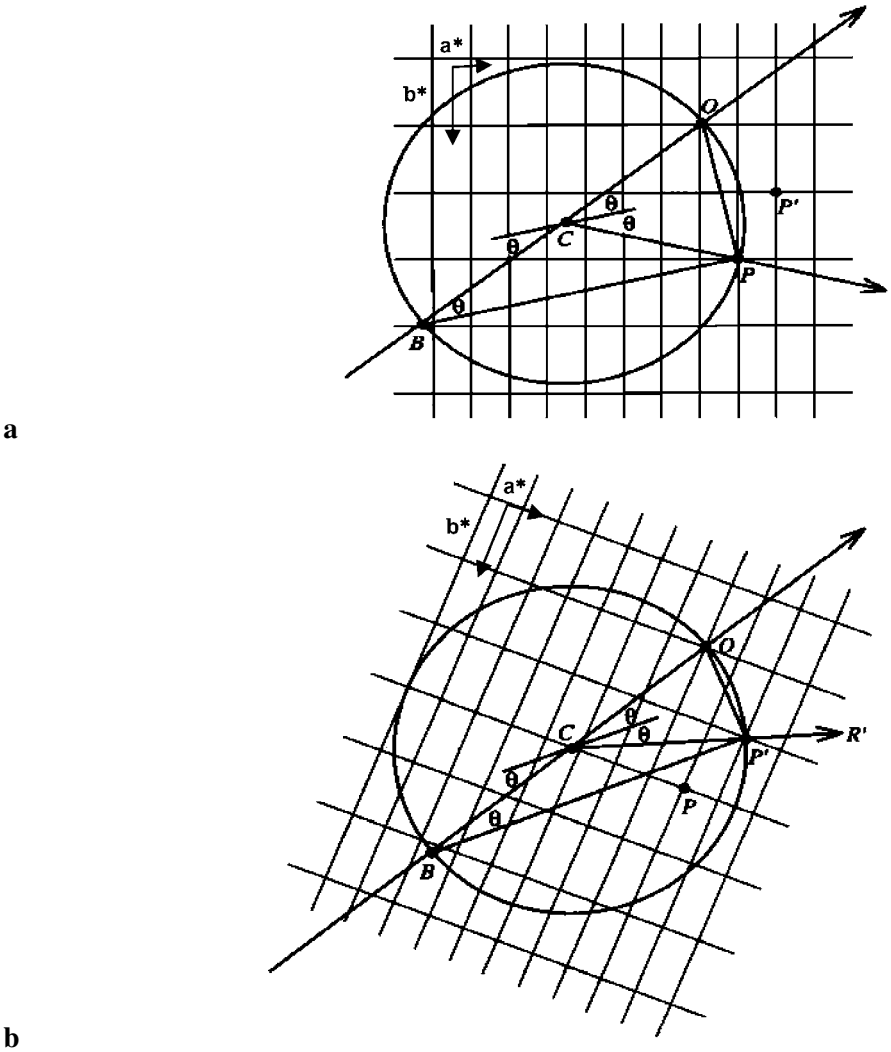
Figure 6. Bragg's Law of reflection. The arrows indicate the electromagnetic waves that are reflected from a pair of parallel planes separated by distance *d*. The dashed line with one end at *B* is a normal to these planes. The bottom wave travels an extra distance, i.e. a path difference of $AB + BC$, and if it is to be in constructive interference with the top wave, $AB + BC$ needs to be an integer multiple of the wavelength λ of the wave. Since the waves are parallel, *OA* and *OC* are perpendicular to *AB* and *BC* respectively, and by simple geometry, $AB = BC = d\sin\theta$. Hence, for the constructive addition of the waves, $AB + BC = 2d\sin\theta = n\lambda$. The amplitude of the reflected wave will depend on the electron density at the point of diffraction.

reciprocal lattice so that a point on the surface of the sphere intersects the origin (O) of the lattice. The vector BCO represents the incident beam. The condition that a particular ray, OP', is a diffracted ray may be expressed as: X-rays will be diffracted in the direction OP' if the point P' represents a reciprocal lattice point ie. the vector P'O is a reciprocal lattice vector $\mathbf{S} = h\mathbf{a}^* + k\mathbf{b}^* + l\mathbf{c}^*$. As the sphere moves about the lattice, a reflection is only observed when the surface hits a reciprocal lattice point.

4. X-Ray Sources

4.1. X-RAY GENERATORS

X-rays are produced when a beam of high energy electrons, which have been accelerated through a voltage in a vacuum, strike a target. In the simplest device called the sealed tube, X-rays can be generated by allowing an electric current to run through a filament that is kept under vacuum in the sealed tube. Electrons



a

b

Figure 7. Reflection in reciprocal space for a two-dimensional case. **(a)** When *P* falls onto the Ewald sphere (depicted by the circle), reflection occurs as both the law of reflection (incident angle is equal to angle of diffraction) and Bragg's Law are satisfied. The line through *C* acts as the plane of reflection. **(b)** As the crystal is rotated, so is the reciprocal space lattice. This brings other reciprocal lattice points, such as P', to the Ewald sphere and equivalently, a new set of planes into a position for reflection. When a reciprocal space vector, OP or OP', falls onto the sphere, the reflection (diffraction) is recorded (after G. Kong).

are made to accelerate, via an applied voltage, from the filament towards a target, usually made of copper or molybdenum, causing X-rays to be emitted from the target. Only a fraction of the energy of the electrons is converted into X-rays with the remainder being dissipated as heat. More intense X-rays can be generated if the target or anode is allowed to rotate fast so that the heat caused by the firing of electrons at it is more rapidly dissipated. So-called rotating anode generators are the X-ray source of choice for protein crystallography laboratories.

X-rays can damage crystals through heating effects and/or the formation of free radicals that can transmit their damaging effects through the solvent channels that run through protein crystals. The effects of X-ray damage can be limited by flash-freezing crystals to about 100 K using a nitrogen stream. So-called cryocrystallography is now the norm for most crystallographic projects.

4.2. SYNCHROTRON SOURCES

For an even more intense source of X-rays, protein crystallographers will often travel to a synchrotron facility. In a synchrotron facility electron (or positrons) are accelerated close to the speed of light by a linear accelerator before being injected into a synchrotron ring where the electrons are kept in a circular orbit through the use of high energy magnets. As the electrons circle around the ring they emit electromagnetic energy at a tangent to their orbit and this energy is funneled down beamlines. Optical elements such as monochromators can be used to select wavelengths of interest.

Synchrotron radiation is particularly useful in obtaining diffraction patterns from very small crystals and also for collecting very high quality diffraction data. Unlike home sources, the wavelength of the X-rays can be varied readily which can be very useful in determining the structure of the protein from the diffraction pattern (see below).

4.3. DETECTORS

The original detectors were film but more accurate and less cumbersome methods were subsequently developed. These developments included multi-wire proportional counters, television area detectors and more recently charged coupled devices. In the last decade film has made a comeback in the form of reusable image plates. These plates store X-ray intensities as latent images in the form of color centers. These are metastable states of trapped electrons in a BaFBr:Eu^{2+} phosphor. The stored image can be read out by scanning the plate with a red He-Ne laser light. The resulting blue stimulated luminescence has an intensity proportional to the number of absorbed X-rays.

Because diffraction patterns often possess symmetry due to the space group of the crystals (see Figure 5), symmetry equivalent reflections are measured during the process of data collection. In addition, the same reflection may be measured

more than once if more data is collected than required to compile a unique set. Measuring multiple copies of reflections is very useful in increasing the precision of data collection and can also be used to calculate a measure of the quality of the data set:

$$R_{\text{merge}} = \frac{\sum_{\text{hkl}} \sum_{i=1}^{N} |\bar{I}(\text{hkl}) - I(\text{hkl})_i|}{\sum_{\text{hkl}} \sum_{i=1}^{N} I(\text{hkl})_i} \tag{13}$$

where R_{merge} (sometimes called R_{sym}) is the residual factor, $I(\text{hkl})_i$ is the i'th measurement of reflection with Miller indices h,k,l, and is the mean value of the N equivalent reflections. The R_{merge} value is typically between 3% and 10%. Sources of error include radiation damage to the crystal during data collection, different absorption properties as the protein is rotated in the X-ray beam, measurements from multiple crystals, and errors inherent in the area detector measurements.

5. The Phase Problem

5.1. INTRODUCTION

The structure factor is the Fourier transform of the contents of the unit cell sampled at reciprocal lattice points h, k and l. Because of the wave nature of X-rays, the structure factors have a phase, ϕ, relative to the origin of the unit cell. Remember that the experimentally determined measurements are intensities for each reflection, h, k and l and the structure factor for each reflection can be determined by equation (11). However, structure factors are complex variables:

$$\boldsymbol{F}_{h,k,l} = |F_{h,k,l}| \exp(i\phi) \tag{14}$$

Hence, only the magnitude, and not the phase, can be extracted from the intensity measurements.

The electron density, ρ, and hence the atomic coordinates of a protein molecule can be determined by performing the inverse Fourier transform of equation (3):

$$\rho(xyz) = \frac{1}{V_c} \sum_{\text{h}} \sum_{\text{k}} \sum_{\text{l}} |F(\text{hkl})| e^{i\alpha(\text{hkl})} e^{-2\pi i(\text{h}x+\text{k}y+\text{l}z)} \tag{15}$$

where V_{c} is the volume of the unit cell, $|F(hkl)|$ is the structure factor of a reflection with Miller indices h, k and l, and $\alpha(\text{hkl})$ is the relative phase of the reflection.

The central challenge in determining protein structures from diffraction patterns lies in the ability to overcome the problem of determining phases for each measured reflection. The major methods of solving the phase problem are outlined in the next few sections.

5.2. MULTIPLE ISOMORPHOUS REPLACEMENT

This was the method used to determined the first protein structures and still continues to be a major method for solving protein structures where no similar structures are already known. In this method the diffraction pattern is measured from crystals of the native protein. Other crystals are soaked in heavy atom solutions and diffraction patterns measured from these crystals as well. Heavy atoms are defined as atoms with sufficient enough electrons around them so they cause a measurable change in the diffraction pattern of the native protein. In practice this usually means choosing atoms such as platinum, uranium, lead, gold and the lanthanides. The 'multiple' in multiple isomorphous replacement (MIR) refers to the fact that at least two different heavy atom data sets must be measured for the method to work. The 'isomorphous' in MIR refers to the fact that ideally the only difference between the diffraction patterns of the native and heavy atom-soaked crystals should be due the heavy atoms. Thus the heavy atoms should bind to the protein in an isomorphous fashion so that they don't disturb any atoms of the protein. The 'replacement' in MIR is a misnomer and a better description would be 'addition'.

The choice of heavy atoms is an art within itself. There are generally two types of heavy atoms: those that are soft and polarisable and which form covalent bonds with protein ligands (examples include mercury, platinum and gold) and those that tend to bind to hard ligands and form ionic interactions (examples include the uranyls and lathanides). The chemical reactivity of heavy atoms can be modified by judicious choice of ligands (e.g. chloride ligands are readily displaced whereas cyanide ligands will remain bound), soak times, heavy atom concentrations, pH and temperature. The most popular heavy atom family are the mercurials which bind covalently to accessible cysteine residues of proteins.

The typical difference between diffraction patterns of the native and heavy atom derivative is between 10 and 30%. Smaller differences might cause difficulty in locating heavy atom positions and the resulting phases will have large errors associated with them. On the other hand very large differences could be a sign of non-isomorphism.

Heavy atom positions in the unit cell are most commonly located using the Patterson synthesis. The general formulation of the Patterson function is a map made from the summation of a Fourier series that has the square of the structure factor amplitudes as coefficients:

$$\mathrm{P}(uvw) = \frac{1}{V_c} \sum_{\mathrm{h}} \sum_{\mathrm{k}} \sum_{\mathrm{l}} F(\mathrm{hkl})^2 e^{-2\pi i(\mathrm{h}x+\mathrm{k}y+\mathrm{l}z)} \tag{16}$$

where u, v and w represent grid units in the Patterson map. A Patterson map is a vector map where peaks represent vectors between heavy atoms (Figure 8).

Protein phases are readily estimated once the heavy atom positions have been located. With reference to Figure 9, vector addition of the scattering factors for individual atoms will give the overall structure factor $\mathbf{F_P}$. Because heavy atoms are electron-rich, their vectors are much longer than vectors due to lighter atoms.

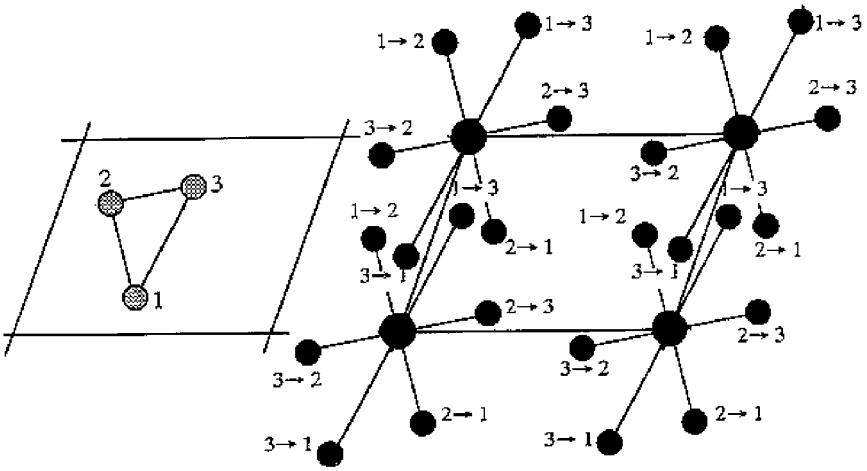


Figure 8. Patterson maps. On the left hand side is a molecule placed in a two dimensional unit cell. On the right hand side is the vector or Patterson map of the molecule.

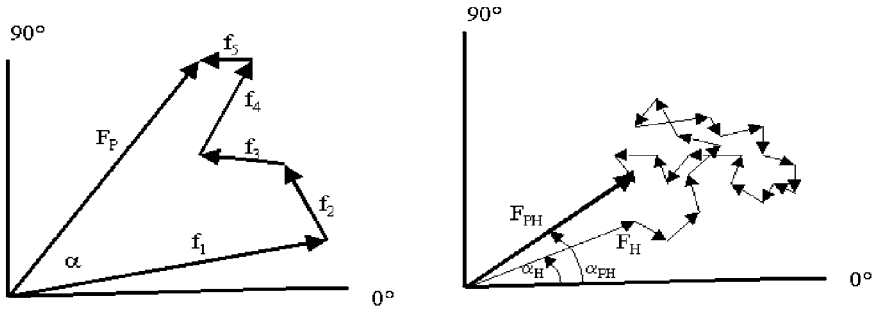


Figure 9. Vector representation of structure factors. On the left hand side the atomic scattering factors of each atom add up to give the overall structure factor, F_P (P is for protein), for a particular reflection with phase α. On the right hand side a heavy atom structure factor, F_H, with its phase α_H, is shown.

Because there are many small vectors due to light atoms in a protein, their vector addition follows a short-stepped random walk. Hence there is a reasonable probability that the angular difference between $\mathbf{F}_H$ (H is heavy atom) and $\mathbf{F}_{PH}$ (PH is protein plus heavy atom) is small and even higher probability that the heavy atom phase (calculated via equation (15)) and protein plus heavy atom phase lie in the same quadrant. Thus the heavy atom phase may be used as a first approximation to the true phase, α_{PH}.

Figure 10 shows vector constructions, called Harker constructions, which are helpful in explaining the heavy atom method. A circle of radius F_{PH} is drawn with center $-\mathbf{F}_H$ so that

$$\mathbf{F}_P = \mathbf{F}_{PH} - \mathbf{F}_H \tag{17}$$

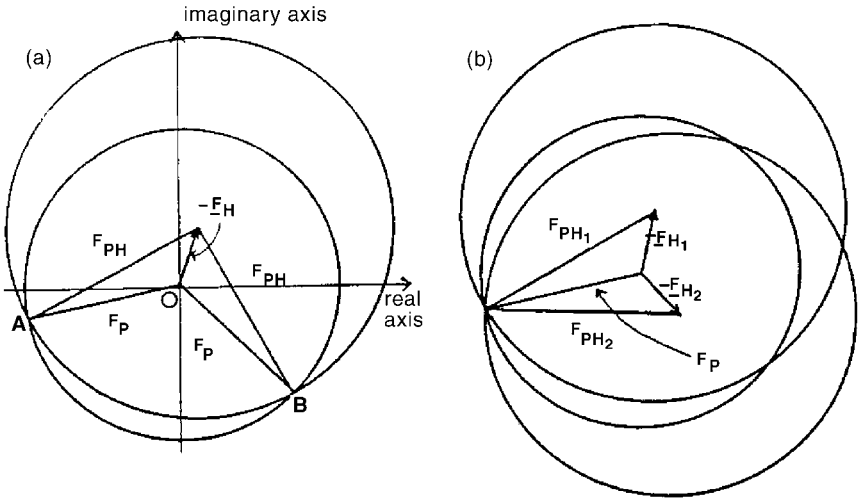


Figure 10. Harker constructions demonstrating the heavy atom method. (a) Phase circles are drawn for native and for a single heavy atom. The circles intersect at two points yielding two possible phase solutions. (b) A unique phase solution is obtained in the case of two independent heavy atom derivatives.

Note that F_{PH} lies somewhere on the circle. We can now draw a circle with radius F_P as in Figure 10. Note that equation (17) only holds where the two circles interact and that there are generally two possible solutions. Hence another piece of phase information is required to find an unique solution. This information may come via a second heavy atom derivative (Figure 10b) or other ways (see below).

The heavy atom method is not without its problems. Poor isomorphism, lack of heavy atom binding and inability to locate of heavy atom sites in the unit cell are commonly met problems.

5.3. MULTI-WAVELENGTH ANOMALOUS SCATTERING

Previously we have assumed that X-rays scatter elastically from the electron clouds around atoms. However, if X-rays interact with more firmly bound inner electrons then there will be a change in energy of the transmitted wave. Anomalous scattering arises when the energy of the incident radiation is close to the resonant frequency of the tightly bound inner shell electrons. The atomic scattering factor of an anomalous scatterer is given as:

$$f_\lambda = f_o + \Delta f'_\lambda + i\Delta f''_\lambda \tag{18}$$

where the $\Delta f'_\lambda$ component is referred to as the dispersion component and $i\Delta f''_\lambda$ component is the absorption or imaginary component because it lags $\pi/2$ behind the primary wave. Close to an absorption edge the dispersion component decreases rapidly whilst the absorption component becomes large (Figure 11). The change

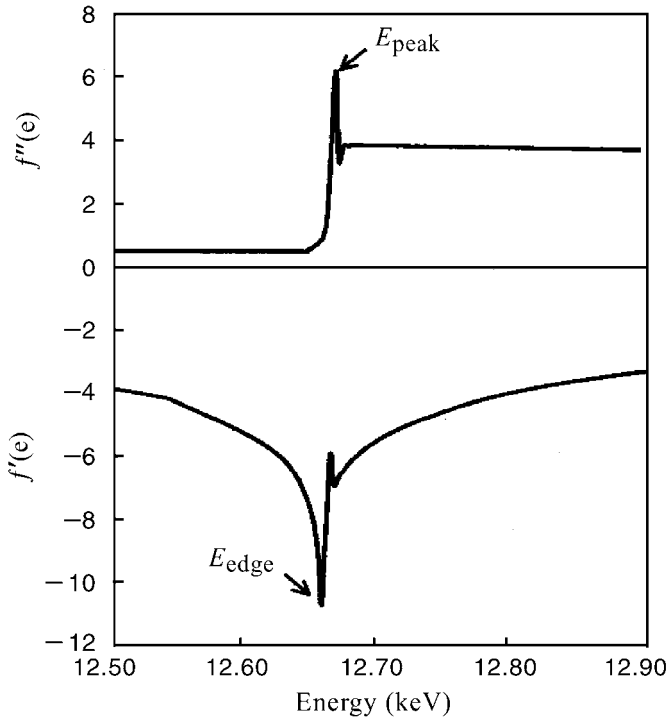


Figure 11. Anomalous scattering factors for a selenomethionine labeled protein. The Se *K* edge occurs at 12.66 eV ($\lambda = 0.98$ Å).

in the diffraction pattern from an anomalous scatterer is usually much smaller (a few percent) than that from the addition of a heavy atom so the data must be measured carefully. Synchrotron radiation has revolutionized the use of anomalous scattering in solving protein structures because of the ability to finely tune wavelengths of X-rays near absorption edges to extract as much signal as possible. Most protein atoms, such as carbon, nitrogen and oxygen, do not have significant anomalous scattering effects in the range of wavelengths that are normally used in X-ray experiments. The most useful atoms are sulfur and metal centers located in metalloproteins. However, by far the most useful element has proved to be selenium which can be incorporated into methionine residues by expressing the protein of interest in the appropriate media enriched for selenomethionine.

In the multi-wavelength anomalous dispersion (MAD) method, data sets are collected from the one crystal at a minimum of three different wavelengths in order to maximize the differences in the real and imaginary components of the anomalous scattering: remote from the absorption edge, at the edge and at the peak. Since the intensities of the diffraction patterns differ it is possible to locate the positions of the anomalous scatters in the unit cell and derive phases from the information using the same techniques used in the heavy atom method.

5.4. MOLECULAR REPLACEMENT

This is conceptually the simplest technique for determining phases and the technique likely to be used to solve most protein structures in the future. If the target protein possesses a similar amino acid sequence (approximately more than 25% pairwise sequence identity) to one for which a structure is already known, then good starting phases can be derived by simply placing the known structure in the correct orientation and position in the unit cell of the unknown protein and calculated phases using equation (15). The correct orientation and position can be determined by placing the probe molecule in the unit cell and calculating its theoretical diffraction pattern using equation (15). The probe molecule is then moved until the experimental and theoretical patterns match. Six dimensional searches (three angles and three translations) are prohibitively expensive to compute so the problem is normally broken up into two searches: a rotation search followed by a translation search. The most common method of searching makes use of Patterson functions so that vector maps are calculated from search and probe molecules and superimposed to find overlaps of vectors.

6. Electron Density Maps

6.1. RESOLUTION

From Bragg's equation (equation (12)), it can be seen that as the scattering angle increases, the separation of reflecting planes is decreasing and hence scattering objects that are close together in space can be resolved. Thus the further the diffraction pattern extends from the position of the incident or primary beam, the higher the resolution of the structure that will be determined (Figure 5). Since X-rays interact with the electron cloud around atoms, the experimentally derived image is in the form of an electron density map. The interpretation of the map is performed using specialist software on a computer graphics workstation. At low resolution (8 to 3.5 Å) the overall shape of the molecule can be seen and helices can be observed as rods of high electron density. At medium resolution (3.5 to 2.5 Å) amino acid sidechains can be identified and the polypeptide chain can be traced. At high resolution (2.5 to 1.0 Å) individual atoms can be located and well-ordered solvent structure around the protein observed. Hydrogen atoms scatter too weakly to be normally observable except at the very highest resolutions.

6.2. DENSITY MODIFICATION

No matter which method is used to solve the phase problem, they all suffer from the fact that the phase values are only estimates and have significant errors associated with them. There are a number of powerful tools available in order to improve the phase estimates and hence the quality of the resultant electron density map. Most of these methods work in real space: the electron density is modified using some

prior knowledge and a new set of phases estimated using equation (15). The most common density modification technique is solvent flattening: in the ideal situation the solvent regions in the unit cell should have uniform electron density but most often these regions are not flat in the initial electron density map due to random noise from errors in the phase and amplitude estimates. In such cases the solvent regions are flattened and new phases calculated. One of the most powerful density modification tools is non-crystallographic symmetry averaging which can be exploited in cases where there is more than one copy of a molecule in the asymmetric unit of the unit cell. In this case equivalent molecules can be identified in the initial electron density map and densities of them averaged to produce a much better set of phases via equation (15).

7. Refinement

The aim of model building is to produce a model that agrees with the experimentally measured diffraction patterns as closely as possibly. This is often expressed in the form of the following equation:

$$R_{\text{factor}} = \frac{\sum_{\text{hkl}} |F_{\text{o}}(\text{hkl}) - F_{\text{c}}(\text{hkl})|}{\sum_{\text{hkl}} F_{\text{o}}(\text{hkl})} \tag{19}$$

where R_{factor} is the conventional residual factor, F_{o} are the experimentally measured structure factors and F_{c} are the structure factors calculated from the latest model via equation (15). Models can be improved by varying the position, x, y and z, and mobility, B, of each atom in order to minimize the residual by least squares methods. Because the number of observations (i.e. the structure factors) are barely more than the number of refinable parameters (atomic positions and temperature factor), then the least squares refinement is poorly determined. This problem is overcome by the incorporation of stereochemical restraints such as bond lengths and angles that are known quite accurately from measurements of small molecule crystal structures.

The non-linear nature of the least squares refinement means that a number of rounds of model building and refinement are required in order to achieve the radius of convergence. A very powerful revolution in the refinement of crystal structures has been the use of restrained molecular dynamics refinement where individual atoms are moved according to Newton's laws of motion using high temperatures of, for example, 3000 K. This has the effect of overcoming atoms being trapped in local minima and hence speeds up the model building/refinement process. The energy function used in crystallographic refinement is:

$$\text{E}_{\text{total}} = \text{E}_{\text{empirical}} + \text{E}_{\text{effective}} \tag{20}$$

The $\text{E}_{\text{empirical}}$ term in equation (13) implements geometric and other restraints:

$$E_{empirical} = E_{bond} + E_{angle} + E_{dihedral} + E_{improper} + E_{vdw} + \ldots \quad (21)$$

In the above equation, E_{bond}, E_{angle}, $E_{dihedral}$, $E_{improper}$, and E_{vdw} represent energies resulting from deviations from ideal bond lengths, bond angles, dihedral angles (for optimizing torsion angles), improper angles (for optimizing planar groups such as benzene rings and peptide bonds), and van der Waals contacts respectively. For example, the bond energy is calculated as:

$$E_{bond} = \sum_{bonds} k_b(r - r_o)^2 \quad (22)$$

where r is the bond length, r_o is the ideal bond length and k_b is the energy constant. The 'ideal' geometry parameters are derived from observations of small molecule structures of amino acids. The $E_{effective}$ component in the overall energy equation (20) represents pseudo-energy terms calculated from structural information such as the reflection data, non-crystallographic symmetry restraints, etc. The effective energy derived from the difference between F_o and F_c is represented by E_{xref}.

$$E_{xref} = W_A \sum_{hkl} w(hkl)[|F_o(hkl)| - k|F_c(hkl)|]^2 \quad (23)$$

where W_A is an overall weight, k is an overall scale factor, and w(hkl) is a weight applied to individual reflections. Thus the crystallography discrepancy term (i.e. numerator of equation (19)) is incorporated into molecular dynamics refinement as a pseudoenergy term.

8. Structure Validation

The typical conventional R-factor for protein models is between 15% and 25% which can be compared to values of less than 5% for small molecule crystal structures. Why the difference? Firstly, the upper resolution limit of the diffraction pattern from protein crystals is nearly always poorer than those of small molecules because of poorer quality crystals and poorer signal-to-noise in the diffraction pattern. Hence many features of a protein structure are poorly modeled. For example, it is rare to see more than the first shell of ordered water molecules around a protein molecule and the solvent regions are crudely modeled by simple mathematical equations. In nearly all cases the temperature factors of each atom are modeled isotropically rather than anisotropically which would be more realistic. Anisotropic B-factor refinement means more refinable parameters per atom but there are rarely enough observations to be able to perform such refinement. Very high resolution crystal structures of proteins show that a significant proportion of surface side-chains adopt more than one conformational state but this cannot be modeled for most protein structures. The relatively poor agreement between model and experiment can lead to a number of errors that must be looked at carefully as model refinement approaches convergence. For example, since model stereochemical parameters are used as restraints, rather than constraints (i.e. fixed values) in

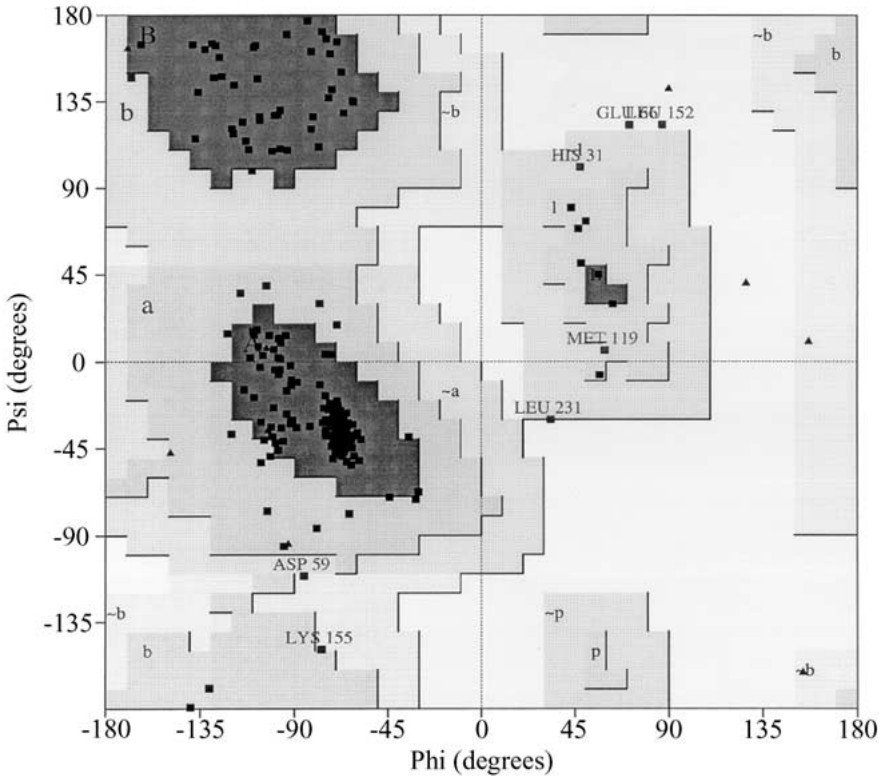


Figure 12. The Ramachandran plot. Favorable areas of the plot are highlighted by the dark gray areas.

the refinement process, they must be checked carefully that they have been allowed to vary within sensible limits. A good model will have bond length deviations less than 0.02 Å and bond angle deviations less than 2.5°. One of the most powerful checks is the Ramachandran or phi-psi plot which is a plot of the main-chain dihedral angles along the polypeptide chain (Figure 12). Because of steric restraints, the conformational space that can be adopted by these angles are restricted. Phi and psi angles are not used as restraints in refinement programs and hence their quality are independent checks of the model quality.

A much more serious issue must be examined for any new protein structure – is the model indeed correct? Unfortunately, there are a handful of published protein structures that have proved to be partially or even totally wrong. The most common causes have been incorrect space group assignment and/or over interpretation of a poor quality electron density map. In one case a protein was actually built back-the-front so the C-terminus ended up where the N-terminus of the protein should have been. In these sort of cases it is possible that the incorrect model can be refined to what look like sensible conventional *R*-factors. One of the most powerful tools to

guard against incorrect model building is the free *R*-factor. In this method, a small percentage of reflections (usually 5–10%) are quarantined from the refinement process. The so-called free *R*-factor, calculating according to equation (19) using only the quarantined reflections, is monitored throughout the refinement process. Decreases indicate correct model building decisions whereas increases suggest the model is being built wrong. A decrease in the conventional R-factor at the expense of an increase in the free R-factor is diagnostic that the model is being overfitted, ie. the large number of refinable parameters is sufficient to cause an improvement in the agreement between the calculated and experimental diffraction patterns even though the calculated model resembles the correct structure much less. Another powerful tool is one based on the expected three-dimensional environment of each of the twenty different amino acids. For example, charged residues normally are located on the surface of a protein and rarely in the protein core. This method, sometimes called 3D–1D profiles [5], determines how well the environment of each residue in the model agrees with what has been observed for correctly built protein models. Poor fits are diagnostic of problem regions in the model. Another powerful tool for structure validation is the Ramachandran plot: many wrongly built models have tended to exhibit almost random distributions in phi-psi plots. Other checks include sensible location of heavy atom sites, correct handedness of helices and sheet, sensible temperature factor trends (e.g. buried residues should have lower B factors than surface residues) and buried charge groups should have their charges dissipated through salt bridges or multiple hydrogen bonding interactions

9. Applications of Protein Crystallography – Rational Drug Design

In the past, the majority of drug discoveries have been based on astute but serendipitous observations or by large screening programs of synthetic and natural products. Advances in molecular biology and protein crystallography have yielded a much more promising method termed rational or structure-based drug design (SBDD). Decades of research have demonstrated that proteins are the site of action for most drugs and hence are the target for the development of new drugs. In SBDD key proteins are identified (for example, by genetic studies or DNA microarrays), crystallised and their crystal structures determined. Through the use of interactive computer graphics and molecular modeling software, it is possible to design potential drugs on the basis that good inhibitors must possess significant structural and chemical complementarity to their therapeutic target. The SBDD methodology requires an iterative procedure in which compounds are designed, synthesised and crystal structures of the protein-drug complexes are then determined to test the modelling predictions. Successful examples of this approach include the currently used cocktail of HIV protease inhibitors against AIDS [6], thymidylate synthase inhibitors against cancer [7] and neuraminidase inhibitors against influenza [8]. In the case of thymidylate synthase inhibitors, more than 100 enzyme-inhibitor complex structures were solved [9].

10. Applications of Protein Crystallography – Functional Genomics

One of the greatest scientific endeavors, the Human Genome Project, has recently been bought to fruition with an estimate that the human genome encodes about 30,000 different proteins [10]. However, the functions of only about one third of these proteins are known with any certainty. Biologists are now embarking on the next big challenge: to decipher the function of all proteins in the human body. This new endeavor, coined by the term functional genomics, is utilizing a variety of powerful tools including X-ray crystallography. It has been argued by advocates that because the function of a protein is encoded by its three-dimensional structure, then structures will lead to a knowledge of protein function. At a practical level all new protein structures are compared to known structures deposited in the Protein Data Bank [3]. Similarities lead to hypotheses that can then be tested by biological assays. Because of the large number of protein structures that need to be determined from many different genome projects, there is a pressing need to speed up the process of solving structures from the current time of a few months. Such research, commonly referred to as structural genomics, is likely to lead to major technological improvements in the coming years. These include crystallization robots, automatic crystal mounting at synchrotrons and automated interpretation of electron density maps. The eventual hope is that protein structures will be determined in a matter of hours rather than the many months it takes at present.

Acknowledgements

I thank Geoffrey Kong and Bill McKinstry with help with the figures. I also thank the Australian Research Council who have supported my work through the award of a Senior Research Fellowship.

References

1. Bernal, J.D. and Crowfoot, D.: X-ray Photographs of Crystalline Pepsin, *Nature* **794** (1934), 133–134.
2. Kendrew, J.C., Dickerson, R.E., Strandberg, B.E., et al.: Structure of Myoglobin, *Nature* **185** (1960), 442–447.
3. http://www.rscb.org/
4. McPherson, A.: *Crystallization of Biological Macromolecules*, Cold Spring Harbor Laboratory Press, New York, 1999.
5. Luthy, R., Bowie, J.U. and Eisenberg, D.: Assessment of Protein Models with Three-Dimensional Profiles, *Nature* **356** (1992), 83–85.
6. Erickson, J., Neidhart, D.J., VanDrie, J., et al.: Design, Activity, and 2.8 Å Crystal Structure of a C_2 Symmetric Inhibitor complexed to HIV-1 Protease, *Science* **249** (1990), 527–533.
7. Appelt, K., Bacquest, R.J., Bartlett, C.A., et al.: Design of Enzyme Inhibitors using Iterative Protein Crystallographic Analysis, *J. Med. Chem.* **34** (1991), 1925–1934.
8. Von Itzstein, M., Wu, W-Y., Kok, G.B., et al.: Rational Design of Potent Sialidase-Based Inhibitors of Influenza Virus Replication, *Nature* **363** (1993), 418–423.
9. Hodgson, J.: Data-Directed Drug Design, *Bio/Technology* **9** (1991), 19–21.
10. *Nature* **409** (2001), 15 February issue.